TU DORTMUND
Fakultät Statistik
Prof. Dr. Philipp Doebler
Dr. Marlies Hafer
Dr. Rouven Michels
Dr. Crystal Wiedner

Winter Semester 2024/2025

Project I
Descriptive analysis of demographic data

The *International Data Base* (*IDB*) of the *U.S. Census Bureau* contains various demographic data (currently from 1950 to 2100) on all states and regions of our world that are recognized by the US Department of State and have a population of 5000 or more. The sources of the database are information from state institutions, such as censuses, surveys or administrative records, as well as estimates and projections by the U.S. Census Bureau itself.

The dataset in the file `census2004_2024.csv` contains a small extract from the IDB. It includes life expectancy at birth and under age 5 mortality rates for 227 countries from 2004 and 2024. For the exact definitions of these variables see `https://www.census.gov/programs-surveys/international-programs/about/glossary.html`. Both variables are stratified by sex, and the countries are divided geographically into 5 regions and 21 subregions. For further details regarding data collection see `https://www.census.gov/programs-surveys/international-programs/about/idb.html`.

Tasks:

1. Describe the frequency distributions of the variables. Consider also the differences between the sexes and regions.
2. Are the values of the individual variables comparatively homogeneous within the individual subregions and heterogeneous between different subregions? To answer this question, first analyse the variability of the values within the individual subregions and then compare the measures of central tendency of the individual variables between different subregions.

   Hint: For this task, choose one of the 5 regions and provide your answer for this region only.
3. Are there bivariate correlations between the variables?
4. How have the values of the variables changed over the last 20 years, i.e. comparing 2004 with 2024?

For tasks 1–3, consider only the year 2024. This project serves to practice the use of explorative and descriptive methods. Therefore, use appropriate statistical measures and graphical methods for the analysis in all parts of the project. Make sure to provide a reasonable interpretation for your results.

Submission

Submission of the report and the corresponding (executable and commented) program code until *Sunday, November 3, 2024, 11:59 pm*, in Moodle. Please submit by the deadline on time, there will be no buffer time.